DOCUMENT RESUME

ED 127 354 TM 005 445

AUTHOR Ree, Malcolm James
TITLE Development of Statistically Parallel Tests by Analysis of Unique Item Variance.
INSTITUTION Air Force Human Resources Lab., Lackland AFB, Tex. Personnel Research Div.
SPONS AGENCY Air Force Human Resources Lab., Brooks AFB, Texas.
REPORT NO AFHRL-TR-76-41
PUB DATE May 76
NOTE 11p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Comparative Analysis; *Computer Programs; Enlisted Personnel; *Item Analysis; Item Banks; Multiple Regression Analysis; Statistical Analysis; *Test Construction; *Test Reliability; *Test Validity
IDENTIFIERS Air Force; *Variable Selection for Multiple Purpose Prediction; Variance (Statistical); VARSEL

ABSTPACT

A method for developing statistically parallel tests based on the analysis of unique item variance was developed. A test population of 907 basic airmen trainees were required to estimate the angle at which an object in a photograph was viewed, selecting from eight possibilities. A FORTRAN program known as VARSEL was used to rank all the test items by unique variance. The item with the highest unique variance was put into a pool, then correlated against each of the remaining items. Next, the best weighted combination of items in the pool was correlated with each of the remaining items until all the items had been included in the pool, and were in order of uniqueness. Odd-numbered items were then allocated to the first test form, and even-numbered items to the second. The two forms were then compared statistically for reliability and validity, and it was found that the means were equal, the variances were equal, and there was equal correlation with a criterion. In order to demonstrate that the technique's success was not a function of the nature of the angle estimation items, a replication on a right-wrong scored attitude measure was done, with similar results. (BW)

AFHRL-TR-76-41

TM

AIR FORCE HUMAN RESOURCES LABORATORY

DEVELOPMENT OF STATISTICALLY PARALLEL TESTS BY ANALYSIS OF UNIQUE ITEM VARIANCE

By

Malcolm James Ree

PERSONNEL RESEARCH DIVISION
Lackland Air Force Base, Texas 78236

May 1976

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Personnel Research Division, Air Force Human Resources Laboratory, Lackland Air Force Base, Texas 78236, under project 7719, with HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved.

LELAND D. BROKAW, Technical Director
Personnel Research Division

Approved for publication.

DAN D. FULGHAM, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
AFHRL-TR-76-41

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
DEVELOPMENT OF STATISTICALLY PARALLEL TESTS BY ANALYSIS OF UNIQUE ITEM VARIANCE

5. TYPE OF REPORT & PERIOD COVERED
Technical Memo

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Malcolm James Ree

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Personnel Research Division
Air Force Human Resources Laboratory
Lackland Air Force Base, Texas 78236

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
62703F
77191221

11. CONTROLLING OFFICE NAME AND ADDRESS
HQ Air Force Human Resources Laboratory (AFSC)
Brooks Air Force Base, Texas 78235

12. REPORT DATE
May 1976

13. NUMBER OF PAGES
12

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)
Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
item analysis
iteration
least squares
multiple regression
reliability
test construction
unique variance
validity

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A method for developing statistically parallel tests based on the analysis of unique item variance was developed. Unique item variance is determined by the multiple correlation of any item with a set of items in an iterative approach.

The approach was successfully implemented on a perceptual ability test and replicated on an attitude scale.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

PREFACE

This research was conducted under Project 7719, Air Force Development of Selection, Assignment, Performance Evaluation, Retention, and Utilization Devices; Task 771912, Armed Forces Programs.

The professional and technical assistance provided during the course of this research by the Computational Sciences Division, and the Testing Branch of the Personnel Research Division is greatly appreciated.

TABLE OF CONTENTS

LIST OF TABLES

DEVELOPMENT OF STATISTICALLY PARALLEL TESTS BY ANALYSIS OF UNIQUE ITEM VARIANCE

I. THE PROBLEM

Most problems of testing concern the validity or the reliability of measurement. Validity is usually determined by the correlation coefficient although other methods are used from time to time. Reliability, on the other hand, is estimated in a variety of ways, for example, split half, alternate forms, or test-retest. The parallel forms reliability coefficient has many advantages. It is the reliability coefficient which accurately shows the proportion of true score variance to total variance. Few assumptions are required; only that the test forms measure the same factor or factors. The parallel forms coefficient is a Pearson product moment correlation and may be tested as any other Pearson product moment correlation. Further, the standard error of measurement may be computed directly from the distribution of difference scores based on the two forms.

Parallel forms reliability is often not computed because it requires at least two forms which have equal means, equal variance, and equal correlation with a criterion. Logically, the items in both forms must sample the same universe; thus a mathematics test and a spelling test, regardless of equal means, variance, and criterion correlation, cannot be parallel forms.

The difficulty usually associated with estimating reliability by parallel forms is building the forms from pools of items. Analysis of unique item variances is a procedure which allows parallel forms to be established from a limited number of items. The procedure for this was developed in order to produce two parallel forms from 120 right-wrong scored perceptual items. This new procedure was developed because traditional methods of producing parallel forms, such as random assignment of items, had failed several times to yield rigorously parallel forms from these items.

II. METHOD

Subjects

A group of 907 basic airmen trainees at Lackland AFB, Texas was selected on a random basis. The pool of 120 items was administered to the subjects in one continuous time period.

These subjects were randomly assigned to a developmental group of 350, a reliability group of 350, and a validity group of 207. Therefore, in each case, cross-validated test forms were tested.

Items

The items were very similar in form. Each required the subject to estimate the angle at which an object in a photograph was viewed. The subject was asked to select the proper angle from among eight angles presented. The same set of eight angles from 0° to 90° served as the answers for each item. Past experience indicated that these angle estimation items formed a homogeneous group (Davis, 1957).

Criterion Measure

The validity criterion was a computer driven and scored free-angle estimation task. The subject was required to push buttons on a keyboard which corresponded to his estimate of the angle-of-view of a series of line-drawn figures. Scores on this task were the algebraic sum of errors for each of ten items. These scores were subtracted from a constant for ease of interpretation and then converted to a unit normal Z score.

Procedure

Ranking Items by Uniqueness. A FORTRAN program known as VARSEL was used to rank all the items by unique variance. VARSEL is described in detail elsewhere (Gould & Christal, 1975), but a brief description is offered here.

VARSEL first computes all the R^2 of each item versus all the remaining items; for example, 1 versus 2 through 120, 2 versus 1 and 3 through 120, 3 versus 1, 2, 4 through 120, etc.

Then the unique or unaccounted-for-variance is computed by

$$U^2 = 1 - R^2 \qquad (1)$$

where U^2 is the unique variance.

The item with the highest unique variance is put into a pool *P*. This item is then correlated against each of the remaining items. From among these resulting U^2's the item with the highest U^2 value is added to the pool *P*.

Next, the best weighted combination of items in the pool is correlated with each of the remaining items, and the highest value of U^2 determined. The item which yields the highest U^2 then becomes the next member of *P*. This procedure is repeated until all of the items have been included in *P*; however, the procedure can be made to stop on any iterative step.

The items are now ranked for uniqueness by order of inclusion from most unique variance to lowest unique variance.

Assigning Items to Tests. Logically, test forms which are statistically parallel should have a high common variance between the scores. This is usually measured by the squared correlation of the scores of the two parallel forms. The task is to allocate items to forms so that no form has too much unique variance.

To allocate the items, it was reasoned that assigning the items selected for the pool on odd-numbered iterations to one form and the items selected for the pool on even-numbered iterations would apportion the unique and common variance equally. The two forms were created in this manner and scored for the reliability and validity groups by the sum of the rights-only method.

Statistical Approach

In order to investigate the effects of the procedure on reliability and validity, the following tests were made:

Reliability. Parallel forms reliability was estimated by computing the correlation between the raw scores on the two forms. This was tested by the following hypothesis:

a. HO: $R_{AA'} = 0$

H1: $R_{AA'} > 0$

The means and variances of the two forms were tested to determine if they differed between forms by the following hypotheses:

b. HO: $\overline{X}_A = \overline{X}_{A'}$

H1: $\overline{X}_A \neq \overline{X}_{A'}$ and

c. HO: $V_A = V_{A'}$

H1: $V_A \neq V_{A'}$

Hypotheses a and b were tested by forming a t-ratio, and hypothesis c was tested by forming an F-ratio.

Validity. The scores on the two forms for the members of the validity group (N = 207) were correlated with the scores on the criterion. Several hypotheses were of interest. These were

d. HO: $R_{C.A} = 0$

H1: $R_{C.A} > 0$

where C is the criterion score and A is test form A,

e. HO: $R_{C.A'} = 0$

H1: $R_{C.A'} > 0$

where A′ is test form A′ and finally,

f. HO: $R_{C.A} = R_{C.A'}$

H1: $R_{C.A} \neq R_{C.A'}$

These hypotheses were each tested by forming a t-ratio. It can immediately be seen that failure to reject the null hypothesis from b, c, and f is the usual test to determine if two forms are parallel. Hypothesis a, d, and e were tested prior to b, c, and f to preserve logic. These five hypotheses were considered as a group, and Type I error was taken for the entire group at $P < .01$.

III. RESULTS

Reliability

Table 1 presents the means, variances, and other descriptive statistics for the two parallel forms for the reliability group.

Table 1. **Descriptive Statistics for Forms A and A′ for the Reliability Group (N = 350)**

Descriptors	Form A	Form A′
Number of items	60	60
Mean	24.71	25.20
Standard Error of the Mean	.44	.46
Standard Deviation	8.15	8.58
Variance	66.44	73.57
Coefficient Alpha	.81	.83

The t-ratio which was constructed to test the difference between the means was .893 (df = 348). Thus, the null hypothesis could not be rejected.

The F-ratio which was constructed to test the equality of the variances was 1.107 (df = 348). The null hypothesis was again not rejected.

The parallel forms reliability coefficient of .825 was tested to determine if it was significantly greater than zero. The t-ratio formed permitted the null hypothesis to be rejected.

Validity

In order for the two forms to be considered parallel, each had to correlate equally with a criterion. Form 1 was found to correlate with the criterion at .2426, and form 2 was found to correlate with the criterion at .2418.

Each was tested to determine if it was significantly greater than zero. The t-ratios for form 1 and form 2 were, respectively, 3.579 and 3.517 (df = 205). Both correlations were found to be significantly greater than zero.

A t-ratio was also computed to determine if these two correlations differed from each other (McNemar, 1949, p. 125). The obtained value was 1.513 (df = 204); thus, the hypothesis could not be rejected.

IV. DISCUSSION

The tests of the hypotheses indicated that the two forms of the Angle-Estimation Test were indeed parallel. The means were equal, the variances were equal, and there was equal correlation with a criterion.

The finding of statistical parallelism is insufficient for determining test forms truly parallel. The one assumption in parallel forms estimation of reliability is that the tests measure the same attribute or combination of attributes. It is surely illogical to impute parallel forms status to two tests that measure different factors, no matter how statistically interchangeable the scores may be. Kelley (1942) presents a fine discussion on this subject.

It was not difficult to make the assumption that all the items in the group of 120 were measures of the same attribute. The items were all of the same type and form, specifically, pictures of models. The task was the same for each item and the possible answer set was the same eight angles. The test constructors (Davis, 1957; Fruchter, 1975) specified that these items were created to measure a single ability.

It was reasoned that if two-parallel forms could be produced by an odd-even split, then further splitting of the scales could produce additional shorter parallel forms. Eight 15-item forms were produced in this manner. Table 2 presents descriptive statistics for the eight 15-item scales. Neither the means nor the variances differ between forms.

Table 2. **Descriptive Statistics of the Eight Scales**

Scale	Mean	Standard Error of the Mean	Variance
1	5.82	.18	6.61
2	5.28	.17	5.80
3	5.84	.17	6.23
4	5.86	.18	7.09
5	6.08	.18	6.79
6	6.07	.18	7.12
7	6.28	.19	7.27
8	5.69	.18	6.79

A factor analysis of the intercorrelation was carried out using the method of Principal components and the Varimax rotation. Only one Eigen value of greater than 1.0 was found. The one factor accounted for 64.8 percent of the variance. Table 3 shows the rotated factor loadings of subtests.

Table 3. **Rotated Factor Loadings of the Subtests**

Variable	Factor I
1	.84
2	.77
3	.79
4	.85
5	.79
6	.80
7	.81
8	.79

The loadings for each scale were all very similar. This indicated that each of the scales was measuring the same factor. This was further evidence that the VARSEL procedure could be used to allocate variance to scales in a manner that produced statistically parallel test forms.

In order to demonstrate that the technique's success was not a function of the nature of the angle estimation items, a replication on a right-wrong scored attitude measure was done. A supervisor's rating served as the validity criterion. Table 4 presents the results of this replication.

Table 4. **Results of Replication on Scored Attitude Scale (N = 259)**

Descriptors	Form A	Form B
Number of Items	23	23
Mean	8.58	7.53
Variance	16.82	16.77
Validity	.21	.26

Again, the five hypotheses were tested with a group Type I error rate of $P < .01$. The tests were found to be statistically parallel.

The results of the replication added evidence to the utility of the procedure.

V. CONCLUSIONS

The procedure of producing statistically parallel test forms by analysis of unique item variance is successful. When the need arises for parallel forms, they can be produced from pools of items which are logically homogeneous.

REFERENCES

Davis, F.B. *Final report–Air Force Contract AF 41(657)-66*, 1957. Bronxville, N.Y.

Fruchter, B. Personal correspondence, 1975.

Gould, R.B., & Christal, R.E. *VARSEL: Variable selection for multiple-purpose prediction systems without using external criteria.* Paper presented at the annual meeting of the Military Testing Association, Indianapolis, IN, September, 1975.

Kelley, T. The reliability coefficient. *Psychometrica*, 1942, 7, 75–83.

McNemar, Q. *Psychological statistics.* New York: John Wiley and Sons, Inc., 1949.

☆U.S. GOVERNMENT PRINTING OFFICE: 1976-671-613/34